El estudio de los aspectos teóricos y descriptivos de lo que se conoce como industrias de la lengua registra un auge importante en Europa desde hace pocos años. Pronto, las lenguas que no se industrialicen dejarán de ser lenguas vehiculares, lenguas de civilización. En este libro se dan a conocer, en el marco de las iniciativas del Consejo de Europa y de la Comunidad Económica Europea, las aplicaciones al uso industrial del desarrollo de la lingüística computacional y el estado actual de la informatización aplicada al lenguaje.

FUNDACIÓN GERMAN SANCHEZ RUIPEREZ

ISBN 84-86168-55-4

9 788486 168551

Las industrias de la lengua

FUNDACIÓN GERMAN SANCHEZ RUIPEREZ

Biblioteca del Libro

Las industrias de la lengua

Bajo la dirección de
José Vidal Beneyto

Biblioteca del libro

Las industrias de la lengua

Las industrias de la lengua

Bajo la dirección de

José Vidal Beneyto

Equipo de traductores coordinado
por MANUEL ALVAR EZQUERRA:
María José Blanco Rodríguez
Francisco Díaz Montesinos
Enrique Lavín Camacho
Manuel Fernando Pérez Lagos
Juan Villena Ponsoda
María Dolores Zamora Navas

FUNDACIÓN
GERMÁN SÁNCHEZ RUIPÉREZ

La Fundación Germán Sánchez Ruipérez es una institución sin fines de lucro, cuyo objetivo general es la creación, fomento y desarrollo de todo tipo de actividades culturales.
Una de sus actividades específicas es la acción editorial, en la que se enmarca la colección
BIBLIOTECA DEL LIBRO.
Sus títulos se editan conjuntamente por la Fundación y Ediciones Pirámide, S. A.

Las INDUSTRIAS de la lengua / bajo la dirección de José Vidal Beneyto ; traducido por Manuel Alvar Ezquerra... [et al.]. - Salamanca ; Madrid : Fundación Germán Sánchez Ruipérez ; Madrid : Pirámide, 1991
496 p. ; 22 cm. - (Biblioteca del Libro ; V)

ISBN 84-86168-55-4 (F.G.S.R.)
ISBN 84-368-0546-1 (Pirámide). - D.L. M-635-1991
1. Lingüística 2. Informática I. Vidal Beneyto, José, dir. II Alvar Ezquerra, Manuel, trad. III. Título
801:681.3

Diseño de la cubierta: Alberto Corazón. Los nuevos medios posibilitan y condicionan nuevos diseños de letras. En la cubierta de este libro se reproduce un estilo utilizado en luminosos de tubo neón.

ISBN: 84-86168-55-4 (Fundación Germán Sánchez Ruipérez)
ISBN: 84-368-0546-1 (Ediciones Pirámide)
Depósito legal: M. 635-1991
Printed in Spain
Impreso en Lavel
Los Llanos, nave 6. Humanes (Madrid)

Índice

Introducción: La industrialización de las lenguas

JOSÉ VIDAL BENEYTO

Razones de un nacimiento

Este libro, y la intervención en él de alguien como yo, ajeno, profesionalmente, al sector científico que representa, es consecuencia de una determinación institucional y de una opción personal. La primera se refiere al programa del Comité Director de la Cooperación Cultural —anagrama CDCC— del Consejo de Europa y a la responsabilidad del Departamento de la Educación, la Cultura y el Deporte, cuya dirección asumo desde hace cinco años, de realizarlo. Ese programa tiene dos grandes objetivos, uno de los cuales es la promoción de la común identidad europea, y, por ende, la defensa y desarrollo de su múltiple patrimonio lingüístico. El desafío cultural que representa la reivindicación de esa identidad y la protección de ese patrimonio en el actual proceso de industrialización de las lenguas europeas, imponían al Consejo de Europa una activa presencia en el mismo. Presencia que puse en marcha y he tenido que pilotar.

La segunda tiene que ver con mi permanente fascinación por cuanto sucede en la periferia de las lenguas, sobre todo de su uso, esos espacios-encrucijada abiertos a lo imprevisto, en los que el lenguaje se topa, a veces, con aconteceres y prácticas de decurso desconocido y se originan interacciones, cruces, rechazos, convergencias, que no se sabe muy bien a dónde pueden llevar. Me sucedió, en los años setenta, con la aplicación del análisis estructural al estudio del fenómeno literario y de la realidad social. Me volvió a suceder en los ochenta con el interface lengua e informática. Lo que es comprensible en mi caso si pensamos que las máqui-

nas, gracias a la informática, querían comenzar, quizá ya habían comenzado, a hablar. La exploración de ese apasionante intento, sus porqués y sus cómos se convirtió para mí en tema de primer interés. Quiero añadir que mi incursión actual no entraña voluntad alguna de ocupar en permanencia un territorio que pertenece a otros.

Es hoy una banalidad recordar que el ordenador no sólo ha modificado sustancialmente nuestra capacidad de almacenamiento y manejo de todo tipo de datos e informaciones, sino que ha ensanchado, de forma insospechada, la reflexión y el análisis de los procesos cognitivos. La simulación en máquina de un espectro, cada vez más amplio, de comportamientos intelectivos se está revelando como un instrumento muy valioso para la exploración de la representación lógica de conocimientos, para el estudio de los mecanismos de aprendizaje y para la modelización de procesos lingüísticos. Ámbito triple e indisociable, tierra de elección de la psicología, la lógica y la lingüística más en vanguardia y que hoy, la informática y la inteligencia artificial están acometiendo con notable éxito.

Y así las lenguas, vehículo privilegiado de comunicación, verbal y escrita entre los hombres, que eran hasta ahora monopolio humano, han comenzado a ser utilizadas por las máquinas y la conversación hombre-máquina y máquina-máquina es una realidad incipiente y limitada, pero efectiva. El mundo mecánico se nos está poblando de voces. La voz del coche, la voz de la cocina eléctrica, la voz del avión de caza, la voz de la cadena de montaje, entre tantas otras, nos informan sobre hechos y nos aconsejan comportamientos que pueden facilitar nuestra vida cotidiana.

Estas proezas verbales que, en muchos casos, suenan más a *gadget* que a verdadero progreso, señalan, de modo llamativo, que el proceso de industrialización de las lenguas está ya en marcha y parece irreversible. Como en tantos otros avatares económicos se trata, también en este caso, de una necesidad hasta ahora mal satisfecha que genera, para su cumplimiento, una demanda potencial. Pensemos por un momento en la invasión de *literatura gris* (informes, actas, resúmenes, partes, certificaciones, órdenes del día, cartas, etc.) que se extiende de día en día y que ni la crisis ni la ola liberal han podido detener. En Francia, por ejemplo, entre las empresas privadas y las administraciones central, regional y para-

pública han superado ya los 400.000 millones de páginas/año. Y esta avasalladora masa textual se produce y difunde mediante procedimientos sólo parcialmente automatizados.

Por no hablar de la traducción, cuyo crecimiento es también exponencial. Dos datos: en la Organización de las Naciones Unidas se traducen más de 300.000 páginas al año; y sólo el manual de pilotaje y mantenimiento del avión *Mirage* exige la traducción «confidencial» de casi 400.000 páginas. Según las evaluaciones más fiables, el mercado mundial de la traducción se acerca a los 200 millones de páginas anuales, equivale a un volumen de negocios superior a los 150.000 millones de pesetas/año y produce más de 180.000 puestos de trabajo.

Era pues inevitable que para este tipo de textos —la traducción literaria es cuestión muy distinta, y su umbral de automatización es muy bajo— se pasase de la práctica individualizada y artesana a comportamientos industriales. De hecho, desde hace más de veinte años, estamos asistiendo a la creación de importantes equipos de traducción instalados en las instituciones y en las empresas —por ejemplo, la Comisión de las Comunidades Europeas cuenta con más de 1.200 traductores permanentes y la sociedad Siemens con casi 200 permanentes y más de 500 temporales— cuyas pautas organizativas responden a criterios de la industria.

Por otra parte, estos equipos utilizan en su trabajo todas las tecnologías de que actualmente disponemos —máquinas de tratamiento de texto, logicales específicos, bancos de datos terminológicos multilingües y, en general, instrumentos informáticos de asistencia a la traducción— que suponen un incremento del 50 al 80 por 100 de su productividad. El mercado que con ello surge, lleva a un importante movimiento industrial. XEROX, gran especialista mundial de burótica, se asocia con SYSTRAN y ofrece un servicio de traducción asistida: ALPS, S. A. y CEGOS ponen en venta un tratamiento de texto multilingüe, adaptado a la traducción, con posibilidad de archivamiento, actualización automática y edición informatizada; NEC anuncia la introducción de un teléfono traductor; IBM vende un logical de interface en lenguaje natural; FUJITSU, HITACHI y TOSHIBA han previsto la comercialización de diversos sistemas de traducción asistida por ordenador. Y tantos otros.

Pero este fecundo desarrollo no debe hacernos olvidar los

límites de la informatización lingüística. Hablar como se hace con frecuencia de traducción totalmente automatizada es referirse a una hipótesis que ni es alcanzable hoy, ni siquiera tiene sentido. Las importantes aplicaciones informáticas en el ámbito lingüístico tienen en sus cuatro principales sectores —el sonido, el léxico, la sintaxis y el sentido— umbrales conocidos que se presentan como infranqueables. Por ejemplo, las variaciones fonéticas de un locutor a otro, e incluso en un mismo locutor, al modificar de forma notable la naturaleza física de los sonidos, confinan la práctica del reconocimiento de la palabra al supuesto de una uniformidad fónica que reduce, considerablemente, sus usos. Por otra parte, el problema del sentido sólo parece abordable, caso por caso, desde una perspectiva coyuntural y empírica, que descalifica la solución global de una modelización generalizadora y convierte la exploración y manejo del hecho semántico en tarea inacabable.

Hablemos pues de lo posible desde la frontera de lo inmediato. En todos los grandes programas tecnocientíficos actuales (ESPRIT, EUREKA, ALVEY en Gran Bretaña, IDS y DARPA en Estados Unidos, los ordenadores de la quinta generación en Japón, los proyectos electrónicos más en vanguardia en Francia y República Federal Alemana) el tratamiento automático de las lenguas naturales es objeto de atención especial y se le asignan objetivos precisos a corto, medio y largo plazo. La generalización de los discos ópticos numéricos (CD-ROM) reforzará aún más estas orientaciones dominantes. Este unánime interés responde a la irrupción de componentes lingüísticos en los más diversos sectores industriales que han multiplicado los posibles usos informáticos de las lenguas y se han traducido en un crecimiento de su mercado efectivo superior al 100 por 100 anual durante los últimos años.

Los correctores ortográficos automáticos incorporados a las máquinas de tratamiento de texto; los instrumentos informáticos de ayuda a la redacción y de generación multilingüe de textos simples (documentos tipificados, correo convencional, etc.); los sistemas de gestión de archivos, información bibliográfica, lexicográfica, terminológica y documental; los mecanismos automáticos de ayuda a la traducción a que nos hemos referido antes; los logicales de análisis lexicométrico o de asistencia en la creación de neologismos; los sistemas informáticos de reconocimiento y síntesis de la palabra, por ejemplo, las máquinas de entrada vocal, aunque sus posibilidades sean (hoy) muy limitadas, constituyen realizaciones industriales que inauguran nuevas prácticas profesionales y nuevos procesos comerciales que parecen susceptibles de transformar radicalmente el paisaje económico y social que heredamos del siglo XIX. Desde esta perspectiva, la llamada revolución de la inteligencia, reclamo publicitario de esta fase del desarrollo tecnológico, convierte a las lenguas naturales —sobre todo a las grandes lenguas de civilización— en materia prima de capital importancia estratégica.

Es urgente que los economistas asuman, con todas sus consecuencias, este hecho capital: la desmaterialización de los procesos económicos hoy más decisivos. Y que los políticos se enfrenten con la nueva condición de los procesos políticos más determinantes: la desterritorialización de su dimensión nacional y/o comunitaria. El territorio que los estados acotan con sus ejércitos y sus fronteras; que la agricultura medía en función de la tierra cultivada y que la industria ceñía a la fábrica, las máquinas y el capital, se extiende y se define hoy por el espacio mental que ocupan los nuevos sistemas y procesos cognitivos.

Por eso, este nuevo ámbito científico e industrial, que representa la interacción de la informática y las lenguas, interpela frontalmente la identidad cultural de pueblos y países y pone en cuestión su misma existencia comunitaria.

Pensemos por un momento en el problema de la aculturación de los logicales. Todo logical comprende, en cuanto a su producción en su país de origen, tres elementos esenciales: 1) las instrucciones en lenguaje informático; 2) los comentarios para los usuarios; 3) los mensajes que se inscribirán en las pantallas. Es evidente que cuando los logicales llegan a manos de sus destinatarios en otro país, están ya *traducidos*, pero la traducción se limita a la parte *mensaje* y deja en su conceptualización original las otras dos partes. Lo que lleva consigo muy importantes consecuencias culturales.

En efecto, dado que todo logical comporta una predicción sobre la base de una representación implícita del comportamiento de los usuarios, para que la traducción supusiera una efectiva *nacionalización* de los logicales sería necesario que se operase una adaptación del logical a los comportamientos propios de los usua-

rios de cada país, en función de las representaciones dominantes en su universo cultural y no en otros que les son ajenos. Sin este proceso de aculturación profunda, acaban interiorizando, sin ser conscientes de ello, modos sociocognitivos que pertenecen a otros universos culturales. No es difícil de imaginar los estragos que ello puede causar en las primeras fases del aprendizaje del niño, apoyadas en didacticales aparentemente neutros.

Las lenguas que no se industrialicen dejarán de ser, en plazo más o menos breve, lenguas vehiculares, lenguas de civilización. No se trata de una apuesta económica, sino de un desafío cultural, de una cuestión centralmente política. Por eso su planteamiento no debe hacerse sólo en términos de competitividad económica sino, sobre todo, de existencia colectiva. Pues si desde el punto de vista de la rentabilidad económica inmediata cabe pensar que la utilización industrial del inglés pueda ser más eficaz y productiva que la diversificación lingüística, es evidente que la supervivencia cultural de un país y de un conjunto de países, de nuestros Estados Unidos de Europa, no es negociable. Si la defensa de la integridad territorial de una comunidad no obedece sólo a consideraciones económicas e incluso, en ocasiones, es injustificable desde ellas, la defensa del patrimonio lingüístico de cada uno de los países europeos y del de todos ellos es también un imperativo metaeconómico. Que Europa no tiene más remedio que hacer suyo.

Realizaciones europeas

Las dos grandes organizaciones europeas que se ocupan activamente del desarrollo de las industrias de la lengua son la Comunidad Económica Europea y el Consejo de Europa. Voy a presentar muy brevemente las principales actividades de ambas en ese ámbito, deteniéndome, algo más, en las del segundo por ser menos conocidas.

La interacción entre lenguas naturales y lenguajes informáticos y los avances en el campo de la ingeniería lingüística se han convertido en factores decisivos para el desarrollo de las tecnologías de la información y de la comunicación que condicionan en nuestras sociedades tanto el crecimiento de la actividad económica como el progreso social. La *Comunidad Económica Europea* no podía quedar ajena a ese proceso y ha lanzado, en consecuencia, un conjunto de acciones, integradas en sus más conocidos programas —ESPRIT, RACE, DELTA, EUROTRA— o formando parte de proyectos específicos de vocación multilingüe.

Previamente al lanzamiento de esas acciones, la Dirección General de telecomunicaciones, industrias de la información e innovación (DG XII) de la Comisión de la CEE promovió una serie de estudios sobre la situación técnica y comercial en los diferentes sectores de la ingeniería lingüística en Europa, en Estados Unidos y en Japón, sobre las necesidades reales de estructuras y recursos, sobre las diversas políticas nacionales, sobre un inventario general de actores, actividades y productos, que sirvieron de apoyo para la elaboración de su plan de trabajo.

Es obvio que la gran variedad lingüística existente en el marco de la Comunidad, base de un mercado potencial importante, y la voluntad institucional de la Comunidad de utilizar en su funcionamiento político y administrativo —en particular en el Consejo Europeo, Consejo de Ministros, Parlamento y Comisión— todas las lenguas de sus estados miembros, hacen del ámbito representado por los países comunitarios, así como de la Organización misma, espacios privilegiados para la experimentación y el uso de los productos y prácticas generados por las industrias de la lengua. El programa EUROTRA es a este respecto una ilustración paradigmática.

Por lo demás, el carácter predominantemente económico de la Comunidad y su especial relación con la industria se traducen en una predilecta atención por los productos y servicios de utilización inmediata, para los que existe ya una demanda efectiva, tales como correctores ortográficos, gramaticales y de estilo; diccionarios electrónicos para la ayuda a la traducción; bases de datos terminológicos; sistemas de ayuda a la edición; aplicaciones CD-ROM; sistemas de búsqueda de textos; lectores ópticos, etc.

El *Consejo de Europa*, por su parte, ha hecho de la multiplicidad de las comunidades culturales existentes en nuestro continente, de la igual consideración intrínseca de cada una de ellas y de su convergencia en un marco común, los elementos fundantes de la identidad europea. Desde ella la reivindicación de multilingüismo aparece como capital, puesto que la lengua es la base y la expre-

sión más relevante de toda comunidad culturalmente diferenciada. Por eso la interpelación que suponen las industrias de la lengua constituye un desafío que no podía dejar sin respuesta.

Consecuente con ello, el Consejo de la Cooperación Cultural reunió en Tours, en febrero de 1986, cerca de quinientos especialistas —lingüistas, informáticos, industriales, escritores, políticos, traductores, representantes del mundo de la economía, etc.— procedentes en su gran mayoría de Europa y de América, para analizar la situación de la lingüística informática y determinar los objetivos prioritarios sobre los que debería concentrarse un programa promotor de la industrialización de las lenguas que respetase, al mismo tiempo, su plena dimensión cultural. En el Manifiesto de Tours que recogió las conclusiones y propuestas de la Conferencia se insta a las instituciones científicas, a las empresas y a los grupos industriales, así como a los gobiernos, para que conjuntamente «continúen el inmenso trabajo que representa el inventario sistemático de las lenguas, comenzado siglos atrás mediante los diccionarios y las gramáticas, extendiéndolo a la dimensión hablada y automatizando su organización».

Para poner en marcha este proyecto se creó en París en mayo de 1986 un grupo informal de coordinación, cuyo cometido principal consiste en promover y animar las actividades de una red europea de cooperación científica y técnica en lingüística informática. Del grupo forman parte directores de centros y laboratorios de Austria, Reino Unido, Francia, Italia, España, Alemania, Suecia, Finlandia, Portugal, Noruega, Yugoslavia y Hungría. A un triple objetivo apuntan las acciones de esta red europea:

a) Compartir e intercambiar, sin restricción alguna, sus saberes, instrumentos y logros; *b*) construir, más allá de sus numerosas diferencias teóricas, una metodología común que permita una efectiva cooperación científica; *c*) hacer que los estudiantes se beneficien de la sinergia así creada, mediante la elaboración de un *curriculum* ideal para la enseñanza de la lingüística informática y de la intensificación de la circulación universitaria entre los diferentes centros.

A la condición informal del grupo de coordinación, se añade su extrema flexibilidad con subgrupos de trabajo que se forman y desaparecen según las necesidades y propósitos de la tarea a realizar. Esta estructura de «geometría variable» consta actual-

mente de los subgrupos siguientes: *corpora* de textos multilingües; informática lingüística fundamental y descripción morfosintáctica de las lenguas; terminología; enseñanzas de doctorado y aplicaciones industriales.

El programa actual fue debatido por el Grupo de coordinación, en la reunión de enero de 1987 en Pisa, y aprobado posteriormente por el CDCC. Sus cuatro grandes objetivos son: la preparación de *corpora* de textos mono y multilingües, la elaboración de léxicos gramaticales electrónicos, la promoción de *corpora* de referencia hablados y el lanzamiento de enseñanzas de postgrado en lingüística informática.

Por lo que se refiere a los *Corpora de referencia escritos*, la reunión de París había puesto de manifiesto la necesidad de disponer de *corpora* de textos sobre el conjunto de las lenguas europeas, dotados de suficiente nivel de homogeneidad para poder realizar estudios comparativos realmente útiles. Pero ello exigía poseer, previamente, *corpora* monolingües de la mayoría de las lenguas europeas, lo que no correspondía en absoluto a la situación actual. Para contribuir a remediar esa carencia se promovieron una serie de iniciativas cuyo objeto era reforzar o iniciar la constitución de esos *corpora*. En dicho sentido, se concluyó un contrato con un equipo de la Universidad de Málaga, dirigido por el profesor Manuel Alvar Ezquerra y apoyado por la sociedad Biblograf del Grupo español Anaya, con el fin de prestarles ayuda en la colecta y repertoriaje de 1.500.000 ocurrencias, destinadas a la preparación de un *corpus* de referencia del español peninsular contemporáneo, con indicación de las frecuencias, el contexto, las formaciones idiomáticas y las construcciones sintácticas. Este *corpus*, cuya ambición es sobrepasar los seis millones de ocurrencias, quiere asumir el idioma en todos sus registros —familiar y cotidiano, literario, científico, técnico, etc.—, para que el resultado estadístico obtenido haga posible un mejor conocimiento de las estructuras del español, tanto desde la perspectiva de la lingüística descriptiva como de la lexicografía. Un acuerdo del mismo tipo con la Universidad de Zagreb permitirá la creación de un *corpus* de referencia de la lengua croata contemporánea, que está comenzando con la recogida y tratamiento informatizado de un millón de ocurrencias, y proseguirá, gracias a la asociación en el proyecto de la Universidad de Belgrado en colaboración con el Instituto

para la lengua alemana de Mannheim, con la producción de un *corpus* bilingüe alemán/serbo-croata.

Los miembros del grupo han querido someter a una primera prueba la compatibilidad de sus *corpora* respectivos y, a dicho fin, han decidido elaborar, a título experimental, un diccionario electrónico multilingüe. Un contrato entre el Consejo de Europa y los centros y/o laboratorios de las universidades de Birmingham, Gotemburgo, Pisa y el citado Instituto de Mannheim, a los que luego se unieron el mencionado equipo español del profesor Alvar y el yugoslavo de la Universidad de Zagreb, ha servido para iniciar la construcción de una metodología común, soporte previo y absolutamente indispensable, para producir diccionarios electrónicos a partir de *corpora* de referencia. La relevancia del proyecto salta a la vista si pensamos en la extraordinaria importancia que reviste la creación de una base de datos lingüísticos, de formato y estructuración comunes, para lenguas tan distintas como el italiano, el sueco, el alemán, el español, el serbo-croata y el inglés.

En relación con los *léxicos-gramáticas*, el Consejo de Europa suscribió sendos contratos con la Universidad de Lisboa, la Universidad Autónoma de Barcelona y la Universidad de Liubliana para el lanzamiento y/o la continuación, según los casos, de léxicos-gramáticas del portugués, español y esloveno, respectivamente basados en la metodología preparada en el Laboratorio de Automática, Documental y Lingüística (LADL) de la Universidad de París VII, bajo la dirección del profesor Maurice Gross, miembro del citado grupo informal del Consejo de Europa. La primera fase de los trabajos, en trance de terminación, se proponía la codificación de 5.000 adjetivos, sustantivos y verbos y, por lo que toca a estos últimos, la confección de un fichero conteniendo las especificaciones de todas las desinencias verbales, de un segundo fichero trabajando sobre el fichero de desinencias, y de un tercero que operase sobre el fichero de conjugaciones. Al mismo tiempo, la elaboración de un programa para decodificar las escrituras alfanuméricas de las entradas verbales del diccionario electrónico permitía realizar, de manera automatizada, la conjugación de dichas formas verbales.

Un segundo contrato del Consejo de Europa con el LADL tenía como objeto estudiar y construir un lenguaje formal unificado para la representación de los sustantivos compuestos. El estu-

dio consistirá en la comparación de las soluciones actualmente utilizadas en determinados ejemplos concretos del francés y de las lenguas latinas, así como en el análisis de la aplicación de estas reglas a las lenguas germánicas (en un primer tiempo el alemán y el danés) referidas a los sustantivos compuestos (sin separador) que tengan una grafía más compleja. En un segundo momento se centrará en la determinación de pautas convencionales suficientemente efectivas para representar los grafos utilizados en la compilación de los léxico-gramáticas electrónicos. Sobre la base de este estudio se presentará un conjunto de recomendaciones para el establecimiento de una norma común que permita la construcción de logicales para la manipulación de léxicos de sustantivos compuestos.

La prioridad de la investigación, tanto fundamental cuanto aplicada, en lingüística informática, es hoy tan imperativa como en febrero de 1986 cuando tuvo lugar el Coloquio de Tours. Es cierto que el volumen de diálogo entre hombres y máquinas ha progresado de manera constante —interrogación de bases de datos, accionamiento de robots, asistencia a minusválidos, traducción, enseñanza y aprendizaje asistidos por ordenador, etc.— pero la industria de la información en su conjunto se ha encontrado frente a un grave estrangulamiento de sus capacidades, derivado de la modesta capacidad actual de los ordenadores para el tratamiento automático del lenguaje natural. Este estrangulamiento, dada la vocación difusora de las tecnologías de la información, reduce de forma notable las posibilidades de innovación tecnológica en ingeniería lingüística, y, como consecuencia, incide negativamente en el crecimiento de las comunicaciones internacionales y en la expansión de los intercambios económicos y culturales.

Si los trabajos realizados en el marco del Consejo de Europa pueden clasificarse, en cuanto a su especificidad temática, en el esquema binario que acabamos de presentar (*corpora* de referencia y léxicos-gramáticas electrónicos), pueden también ser objeto, por lo que se refiere a sus usos o a su condición innovadora, de otro tipo de clasificación, en el que, por una parte, se agrupen *los trabajos de aplicación-perfeccionamiento* de los sistemas productos ya existentes, y, por otra, se reúnan las experiencias tendentes a elaborar métodos y prácticas inventivos y renovadores. Entre los primeros deberíamos situar los ya antes aludidos *corpora* de refe-

rencia monolingües (español, croata), bilingües (alemán/serbo-croata) y léxicos-gramáticas electrónicos monolingües (español, portugués, esloveno). Es evidente que respecto de las realizaciones muy costosas y de gran aliento, la función del Consejo de Europa no puede consistir sino en promoverlas y lanzarlas, buscando luego, por medio de los estados miembros y de sus organizaciones no gubernamentales (ONGs) las instancias más adecuadas —administrativas, científicas, industriales— es decir, los actores sociales y económicos, verdaderos protagonistas de ese ámbito, para sacarlas adelante.

Entre las *actividades de condición innovadora* quiero limitarme a dos: el estudio y producción de un lenguaje formal unificado para la representación de nombres compuestos y la ya señalada elaboración de un diccionario electrónico multilingüe experimental. En el primer caso se trata de construir un modelo unificado de transposición a las lenguas germánicas —comenzando con el alemán y el danés— y eslavas —al principio sólo el esloveno— de un lenguaje formal cuya efectividad ha sido ya probada en las lenguas latinas (añadamos que el ejercicio de transposición se apoya en las experiencias realizadas por el LADL de París en relación con el griego moderno). Por lo que toca al segundo, su objetivo último es, más allá del propósito nada despreciable de fabricar un diccionario electrónico multilingüe, la armonización de los métodos de constitución de *corpora* monolingües, o, en otras palabras, el establecimiento de pautas y principios, científicamente verificados y aceptados por todos, para la producción de *corpora* monolingües. Recordemos que el trabajo en curso afecta al alemán, inglés, español, italiano, sueco, serbo-croata y, muy pronto también, al húngaro.

El grupo de expertos de industrias de la lengua ha insistido en la necesidad de acelerar los trabajos de automatización lexicográfica en las direcciones siguientes: determinación y representación de las palabras compuestas, desambiguación, ubicación, tipicalización, mejora de la efectividad de los lexicales para el análisis gramatical, etc. La superación de los paralizadores estrangulamientos a que acabamos de referirnos sólo será posible si en cada una de estas operaciones, y en la totalidad del ámbito, se consiguen logros suficientes para poder disminuir la intervención humana a que obligan las actuales carencias del proceso automático.

Sólo desde el zócalo formado por el conjunto de normas unificadas y aceptadas por —no impuestas a— la comunidad científica podrá abordarse responsablemente el urgente desafío que representa el tratamiento automático de la palabra. Comenzando con los *corpora de referencia orales*, es decir los que se proponen recoger y tratar automáticamente la palabra, que sirven a un conjunto de intereses concretos: pedagógico (análisis de la adquisición del lenguaje y elaboración de métodos de enseñanza y sobre todo de autoaprendizaje fundados en ese análisis), médico-social (elaboración de técnicas de rehabilitación del lenguaje), industrial (numerosas aplicaciones relacionadas con la síntesis de la palabra), etc. Y sobre todo, al gran interés cultural que representa la salvaguarda y promoción de las lenguas de difusión limitada para las que no tenemos apenas testimonios escritos. En este sentido existen acciones en curso referidas a las grandes lenguas occidentales y a algunas otras de reducida circulación como el frisón y el lapón y están en proyecto acciones de la misma naturaleza sobre el galéico y el rom hablado por las poblaciones nómadas de Europa Central y Oriental. Estos trabajos de condición germinativa hacen augurar la confirmación del tratamiento automático de la palabra en un número importante de lenguas.

Materiales y estructura de este libro

Para que los ordenadores lean, entiendan, interpreten, traduzcan o corrijan el lenguaje natural, escrito o hablado, de los seres humanos ha sido necesario que convergieran y se concertasen saberes y prácticas procedentes de disciplinas y escuelas muy diversas. Podrían presentarse múltiples itinerarios de estos encuentros, desencuentros, rupturas, alianzas entre los lenguajes naturales y las máquinas que darían lugar a otras tantas historias del proceso que constituye la materia de estos textos.

Entre todas esas posibles historias hemos elegico cinco para iniciar esta compilación que tienen como propósito introducirnos en el ámbito del tratamiento automático del lenguaje en su más reciente despliegue. Ninguna de esas cinco historias pretende por sí sola ofrecernos una visión completa de la situación actual ni de los caminos y etapas que han conducido a ella. Ni siquiera consi-

deradas en su conjunto podrían esas cinco contribuciones tener la pretensión de ofrecernos un panorama integral, pues para ese cometido serían necesarias muchas otras vías.

Al contrario, se ha renunciado de forma explícita, y conviene anotarlo en este punto, a la presencia de numerosas materias y disciplinas —lo que hace que se hable muy poco en este libro de, por ejemplo, sociolingüística, filosofía del lenguaje, filología, psicolingüística, etc. —no por infravaloración o desconsideración de esos importantes campos temáticos y científicos sino porque la opción central en la que se encardinan todos los materiales aquí ofrecidos es muy otra, a saber: su dimensión práctica, su capacidad operativa. En otros términos, se han querido privilegiar las contribuciones que tenían alcance aplicativo, que describían objetos, cualquiera que fuese su naturaleza, de la forma más precisa y ejecutiva posible.

Estos objetos pueden ser listas de palabras, aparatos procesadores, formas de resolver un problema lingüístico, instituciones financiadoras, programas de investigación, pero todos tienen en común la condición de ser elementos de una práctica, que, de una manera o de otra, es responsable de la eficacia de los ordenadores susceptibles de operar con palabras y frases. La primera parte «El tratamiento automático del lenguaje natural» y los textos de Gérard Sabah, Henry Thompson, Jean-Sylvain Liénard y Maurice Gross tienen como intención instalarnos globalmente, desde cuatro esquinas operativas distinas, en esa perspectiva en la que la informática tiene un rol preponderante.

En efecto, entre todas las disciplinas científicas que han cooperado en la constitución de las industrias de la lengua, la informática ha tenido, a pesar de su carácter, según muchos instrumental, y de su corta edad, comparada con la de otros sectores científicos, una función decisiva.

Para fundar esta condición matriz de la informática, baste con recordar que el desarrollo de la rama de la lingüística de que nos ocupamos en esta obra no se inicia realmente hasta 1955 gracias a la aparición de los ordenadores. Por lo demás, este fenómeno no es exclusivo del sector lingüístico. El auge de muchos nuevos campos científicos en las disciplinas tradicionales, de la filosofía a la física y de la psicología a la mecánica, es resultado directo de la instalación y uso de ordenadores en todos los laboratorios del

mundo. La informática, al asumir este papel germinal, se sitúa en la esfera del saber, no como un artilugio técnico de vocación meramente auxiliar, sino que, como queda dicho anteriormente, se constituye en fermento iniciador de procesos de crecimiento científico, sin los que regiones enteras del nuevo conocimiento no hubieran sido aún exploradas.

El ámbito informático se divide, a su vez, en numerosos subámbitos que se agrupan habitualmente en torno de dos polos: el material y el logical. Los progresos considerables realizados a lo largo de las dos últimas décadas, en el conjunto de estos subámbitos ha hecho posible que el tratamiento automático de las lenguas naturales comenzase a ser una realidad. Un solo ejemplo será suficiente para ilustrar esta afirmación: la lectura óptica de textos. La operación es conceptualmente muy sencilla: su objeto es transferir un texto impreso en una hoja de papel a la memoria de un ordenador sin que se trate de comprenderlo, ni de interpretarlo ni de corregirlo. Basta con que todo carácter que aparezca impreso en el papel sea leído correctamente, es decir reproducido con fidelidad por la máquina. Para realizar esta misión adecuadamente, es decir mejor y más deprisa que lo haría un operador humano, han sido necesarios cerca de veinte años de esfuerzos para lograr aumentar la capacidad de las memorias electrónicas, acrecer la velocidad de cálculo de los microprocesadores, disminuir el número de componentes electrónicos y su consumo eléctrico, producir pantallas catódicas más poderosas y fiables, disponer de lasers de pequeño tamaño y de motores paso a paso, concebir sistemas de explotación y lenguajes de programación más potentes y finalmente elaborar logicales de reconocimiento de caracteres gráficos susceptibles de utilizar todos esos avances. Si en un sólo subámbito se hubiera producido un solo bloqueo conceptual o tecnológico, la lectura óptica de textos hubiese sufrido un notable retraso y en cualquier caso no existiría todavía hoy. Pues la informática, como cualquier otra actividad científico-positiva, es un proceso esencialmente gradual y acumulativo, en el que las rupturas no cancelan sino que asumen el anterior patrimonio de logros.

Ni era posible ni tenía sentido describir en esta obra el conjunto de subámbitos informáticos que intervienen en el tratamiento automático de las lenguas naturales. Por otra parte, la extrema

especialización de esos subámbitos hace su comprensión difícil y su utilidad limitada para lectores cultivados pero procedentes de otros sectores científicos. Por esta razón se ha optado por no entrar en desarrollos excesivamente técnicos o de detalle y por reducir al mínimo la utilización de ecuaciones, algoritmos, esquemas electrónicos o descripciones de logicales. Sin embargo, nos ha parecido necesario mostrar de qué manera los teóricos de la informática abordan los problemas con que nos enfrentamos al intentar hacer manipular el lenguaje natural por parte de las máquinas. Esta es la razón que justifica la presencia de la diáfana contribución de Maurice Nivat y Dominique Perrin que cierra la primera parte.

Cualquier proceso de reflexión que aspire a traducirse en un conocimiento concreto, como sucede con todo comportamiento científico, comienza por reunir un conjunto de materiales que forman la base común de quienes por él se interesan y quieren participar en él. Las teorías y su validación/invalidación no serían posibles si la confrontación no tuviera ese zócalo común, si no se realizase sobre los mismos objetos y no respondiera a las mismas pautas operativas. En nuestro campo, esa base común está constituida por los *corpora* de referencia que deben construirse con plena independencia de la disciplina o subdisciplina en que se trabaje, de las opciones teóricas o metodológicas de las que se parta, y del destino del producto que de ellos pueda y quiera derivarse.

Pero un *corpus* de referencia no consiste sólo en la agregación de una gran cantidad de datos, sino que exige una estructuración muy precisa que disponga de métodos, algoritmos de análisis e instrumentos de mantenimiento y manipulación. Además, su estructura debe tender a la máxima flexibilidad e incluir una amplia gama de usos practicables a fin de no confinar a los futuros usuarios en las solas posibilidades contempladas en el momento de su elaboración inicial. De aquí que sea capital establecer una metodología común, no sólo a los efectos de la comparación de *corpora*, a que nos hemos referido al hablar de las acciones del Consejo de Europa, sino también para las adaptaciones, transformaciones e incluso el simple acrecentamiento de cada *corpus*.

La segunda parte de la compilación, que lleva por título «La constitución de *corpora* de referencia», responde a esos objetivos y

tarea. Los tres textos de John Sinclair, René Carré y Antonio Zampolli presentan de manera a la vez completa, directa y sencilla, es decir realmente didáctica, los criterios organizadores y las operaciones necesarias para la creación de *corpora*, tanto en lenguaje escrito como hablado.

El tratamiento automático de las lenguas naturales, bien sea en forma de análisis de textos, con el fin de proceder a una indexación o a la interrogación de un banco de datos por ejemplo, bien sea como generación de textos, exige que se disponga, previamente, de una representación sistemática de esas lenguas naturales directamente utilizable por las máquinas de modo automático. Pero a diferencia de los seres humanos que en sus procesos de aprendizaje pueden recibir y asimilar grandes zonas de saber implícito, los ordenadores son sistemas a los que es necesario «decírselo todo». En otras palabras, si las descripciones contenidas en los diccionarios y las gramáticas tradicionales (sobre soporte papel o sobre soporte magnético) pueden ser suficientes para los lectores humanos dotados de una competencia lingüística normal, son patentemente insuficientes —y todos los intentos realizados hasta ahora lo han confirmado— para su utilización informática.

Un léxico-gramática electrónico tiene como destino su utilización por un analizador o un generador automáticos lo que implica que posea informaciones sistemáticas sobre los posibles sujetos y complementos que puedan ofrecerse a cada verbo, así como indicaciones formales sobre las diversas posiciones y formas que puedan asumir los sujetos y los complementos en una frase cualquiera, al igual que sobre las posibilidades de pronominalización, de pasivación, etc. La realización de un léxico-gramática de esta naturaleza supone un minucioso trabajo de análisis sintáctico y de formalización, un verdadero «desmontaje» de la lengua; es decir, la descripción exhaustiva y sistemática de absolutamente todos sus elementos con el objeto de elaborar una representación utilizable por los autómatas. La representación de esa descripción se efectúa mediante tablas a las que deben poder tener acceso los autómatas. Este tipo de tablas constituyen la materia nodal de los diccionarios electrónicos propiamente dichos que nada tienen que ver con las versiones automatizadas de los diccionarios convencionales, aunque ambos tengan soporte magnético.

La constitución de diccionarios electrónicos y de léxicos-gra-

máticas es, pues, pieza fundamental en el tratamiento automático de las lenguas naturales escritas. El quehacer a que esa tarea invita tiene mucho que ver con el de los enciclopedistas del siglo XVIII, en cuanto que ambos postulan una organización de saberes consciente de sus aprioris, o sea de su condición arbitraria, a la vez que una exhaustividad en el propósito, tan grande como los recursos humanos disponibles en ese momento lo permitan. Así considerada, la constitución de diccionarios electrónicos es obviamente una labor, potencialmente, sin fin. Sin embargo, a mediados de los años ochenta, los diccionarios electrónicos disponibles alcanzaron ya un dintel crítico, diríamos suficiente, para permitir su explotación industrial, aunque no cubriesen más que un porcentaje muy exiguo del conjunto de fenómenos lingüísticos, como nos señala Maurice Gross a propósito del francés para las palabras compuestas.

Como se ha indicado anteriormente al resumir las actividades europeas en el campo de las industrias de la lengua, los diccionarios electrónicos y los léxicos-gramáticas son uno de los instrumentos más eficaces para estimular el progreso de la lingüística comparada, a condición de que se realicen por familias de lenguas, se coordinen los distintos desarrollos y se adopten procedimientos idénticos o equivalentes. En este sentido, y sin pretender acercar el agua a mi molino, me parece fundamental promover, a nivel mundial, la constitución de redes que conecten los centros e institutos de lingüística informática por familias de lenguas, con el fin de contribuir a dotar el área de cada lengua del instrumental lingüístico necesario para responder al desafío de su industrialización, sin tener que plegarse a los intereses económicos, por lo demás coherentes y legítimos, de las grandes multinacionales de la información y la comunicación. Y quiero añadir que la experiencia del Consejo de Europa, a este respecto, ha probado, en el ámbito de las lenguas latinas, que el sistema de funcionamiento en redes es el más apropiado para conseguir la asociación efectiva de equipos de nivel técnico desigual en un mismo trabajo.

La parte tercera titulada «La elaboración de diccionarios electrónicos y sus aplicaciones» recoge cinco textos: dos del profesor Zampolli en los que se exponen lo que son y cómo se constituyen los bancos de datos léxicos, en especial de carácter multifuncional, cuáles son sus principios y cuáles sus pautas y modos de operar,

todo ello con rigor pero de forma eminentemente pragmática; uno del profesor Maurice Gross explicando las diferencias entre diccionarios usuales y electrónicos y presentando el funcionamiento del sistema DELA, que privilegia el enfoque morfológico y ha sido elaborado en el Laboratorio de Documentación Automática y Lingüística que él dirige, sistema a partir del cual se han compuesto los diccionarios electrónicos del francés: DELAS, DELAR y DELAF; otro de Laurence Danlos que tiene como núcleo central la generación automática de textos y nos expone el rol de la precodificación de textos, sin y con variables, de los diccionarios electrónicos y de los léxicos-gramáticas en la construcción de sistemas generativos; y finalmente una presentación a cargo del profesor Manuel Alvar Ezquerra sobre los trabajos lexicográficos automatizados del español, que tiene como hilo conductor el relato minucioso y didáctico de la preparación de la edición de 1987 del *Diccionario general ilustrado de la lengua española VOX* (DGILE) y de sus transformaciones y perfeccionamientos posteriores, en especial la construcción de una gramática de ese diccionario, susceptibles de convertirlo en una verdadera base multifuncional para la construcción de otros diccionarios electrónicos del español.

La conceptualización científica y técnica ha tenido en el siglo XX un crecimiento exponencial, y la posibilidad de disponer de palabras capaces de expresar, con la mayor univocidad posible, esos numerosísimos nuevos conceptos se convirtió, ya en los años cuarenta en una necesidad de satisfacción inaplazable. La creación de la Organización Internacional para la Normalización (ISO) respondió a ese imperativo, que no podía encontrar su cumplimiento sino en la formulación de principios comunes y en la propuesta de aplicación de métodos análogos para la producción de nuevos términos. La terminología como área interdisciplinaria, tanto científica como académica, tiene ya un casi unánime reconocimiento. En esta situación, la informatización y automatización del ámbito lingüístico y sus posibles aplicaciones tenía que alcanzar de lleno a las actividades terminológicas y constituirlas en un elemento importante de las industrias de la lengua. Y así ha sido. En la cuarta parte, bajo el enunciado «Los inventarios terminológicos», se recogen un texto de Christian Galinsky y dos notas de Wolfgang Nedobity, que presentan las principales realizaciones,

temas y problemas con que ha tenido que enfrentarse la terminología en su proceso de automatización. El primero, como director del Centro Internacional de Información para la Terminología (INFOTERM), promovido por Unesco y creado en Viena en 1971 con el fin de coordinar los estudios sobre terminología en su dimensión internacional, expone los cometidos actuales de su organización y sobre todo los programas de trabajo de la Red Internacional de Terminología (TERMNET) cuya preparación corre a cargo de Infoterm y que encuentran en la informatización el soporte esencial de sus acciones. Por su parte, Wolfgang Nedobity se ocupa de elucidar las posibilidades de crear una metodología común para el tratamiento de los lenguajes especializados y de la ayuda que quepa esperar a ese respecto de la informática.

La parte quinta, «Ejemplos de cooperación europea», presenta, de la mano del ya mencionado profesor Maurice Gross y del doctor Loll Rolling, de la Comisión de las Comunidades Europeas, algunas de las principales actividades que el Consejo de Europa y la CEE han puesto en marcha en este campo, y a las que acabamos de referirnos en la parte segunda de esta Introducción. El propósito de esta parte no es ofrecer una descripción exhaustiva de todo lo que se está realizando, sino mostrar, con algún detalle, las experiencias más avanzadas y que parecen más innovadoras. La reflexión de Jean-François Degrémont con que concluyen estas contribuciones, aunque sea de carácter más general y su intención claramente epistemológico-metodológica, encuentra su ubicación en esta parte porque la materia sobre la que trabaja está hecha de experiencias y realizaciones europeas.

Las lenguas que hablamos, las lenguas habladas, tienen, desde una perspectiva informática, características muy especiales que hacen que su aprehensión y tratamiento por las máquinas tenga que ser muy distinto del que conviene y funciona en el caso de las lenguas naturales escritas. Para nosotros, los humanoparlantes, lengua escrita y lengua hablada parecen dos formas muy próximas de una misma actividad comunicativa y expresiva. Pero para el investigador que intenta construir modelos informáticos, que sirvan tanto para el reconocimiento como para la generación de formas orales, las diferencias son tan grandes que habría que considerar ambos ámbitos como totalmente diferentes. Pues ni pueden utilizarse en ambos casos los mismos modelos lingüísticos,

ni cabe operar con los mismos métodos de tratamiento, ni es posible utilizar los mismos modelos informáticos. Esta dicotomía deriva principalmente de la necesidad de controlar y de interpretar los fenómenos vibratorios que, por definición, no intervienen en la lengua escrita. Hay que añadir que la historia y el tratamiento automático de la palabra han estado fuertemente condicionados por el hecho de que una industria tan importante como la de las telecomunicaciones se ha interesado, directamente y desde el primer momento, en su desarrollo a causa de su interés por obtener, lo más rápidamente posible, resultados aplicables al proceso industrial. En cualquier caso, y como veremos en la sexta parte del libro, que se ha titulado «La comunicación hablada: reconocimiento y síntesis del habla», las cuatro contribuciones aquí reunidas prueban que el tratamiento automático de la palabra consiguió liberarse pronto de una cierta ganga inicial y desarrollarse en direcciones nuevas y productivas. A ese efecto Gunnar Fant, Björn Granström y Rolf Carlson por una parte y Joseph Mariani, por otra, en sendas aportaciones nos introducen en la problemática informática de la comunicación hablada centrándose, sobre todo el segundo, en los niveles, materiales y sistemas de reconocimiento de la palabra, aunque sin olvidar los aspectos ligados a la operación de síntesis, en particular, los que se refieren al lenguaje comprimido y a los sistemas de síntesis a partir del texto. Por su parte, el profesor René Carré nos informa sobre las actuales condiciones de la investigación en comunicación hablada y su previsible evolución en el inmediato futuro, y el profesor John Laver recoge y analiza los puntos principales del Seminario sobre Tecnología del Lenguaje, organizado en Aarhus en el marco del programa ESPRIT de la CEE, insistiendo, con razón, en los aspectos relativos al lenguaje hablado que representaron uno de los núcleos más interesantes de aquella importante confrontación de expertos.

En la parte séptima, «La enseñanza de la lingüística informática», se aborda el tema de la formación en esta especialidad, porque, al igual que sucede en todos los otros sectores industriales, hoy es imposible tratar de cualquier industria sin acometer, simultáneamente, el proceso de formación de los destinados a trabajar en su ámbito. *Hacer* y *saber hacer* se han convertido en prácticas indisociables y las industrias de la lengua no han tenido más

remedio que asumir esa realidad y plantearse con ello el problema consustancial a todo aprendizaje de un hacer concreto, en especial, de carácter técnico: cómo conciliar la adquisición, por parte de los estudiantes, de un saber general que les capacite para incorporar en el futuro nuevos conocimientos específicos, de cuyos fundamento y coherencia puedan dar razón, con la posibilidad de disponer de inmediato de las competencias precisas y del alto nivel técnico propios de un dominio de tan notable complejidad. El trabajo a realizar en las industrias de la lengua no puede satisfacerse ni con la preparación de teóricos puros ni con la de simples programadores informáticos. El único nivel teórico válido en este campo es el que surge del enfrentamiento de la informática con la materia prima lingüística, al igual que las prácticas técnicas más eficaces son las que derivan de la más amplia y profunda comprehensión informática de los fenómenos lingüísticos.

Por eso, la reivindicación del nivel académico máximo, el doctorado, como punto culminante en las formaciones de este tipo, parece ser el marco más adecuado para dar efectiva respuesta a estas exigencias y expectativas. La existencia en diversos países del título de doctor ingeniero es un precedente que ha dado ya pruebas de su utilidad. Ahora bien, desde una perspectiva europea esta actividad educativa sólo puede acometerse coordinadamente, creando redes de cátedras y de centros empeñados en la misma tarea en distintos países, llevándoles a concertar sus presupuestos teóricos, epistemológicos y metodológicos, conjuntando sus procesos formativos para establecer *curricula* en parte análogos y en parte complementarios, intercambiando en permanencia experiencias, profesores y estudiantes, en una palabra trabajando en un marco común sin renunciar a la autonomía ni a la especificidad de cada cual. Al hablar en esta Introducción de las realizaciones europeas he presentado brevemente el camino andado por el Consejo de Europa en esa dirección. En las tres contribuciones de esta parte, a cargo de Bernard Cassen, Jean-François Degrémont, Helmut Schnelle y Antonio Zampolli, la primera, y de Maurice Gross y de Antonio Zampolli, la segunda y tercera, respectivamente, se relatan algunas de las actividades, ya en ejercicio, de la formación postgrado en lingüística informática, especialmente en Italia (Universidad de Pisa) y en Francia (Universidad de París VII), así como los supuestos, posibles *curricula* y mecanismos universitarios para la creación de la red europea de un doctorado común en esa nueva disciplina.

En la octava y última parte de esta compilación, «Muestras de una industria», se relatan algunas experiencias que me han parecido significativas del desarrollo industrial del nuevo ámbito. El criterio que ha guiado su selección es el mismo que ha orientado la construcción de este libro: considerar que las industrias de la lengua no son un sector sólidamente establecido y definitivamente organizado sino en fase de gestación y que su condición emergente, su naturaleza de proceso *in fieri*, con aciertos y errores, pistas falsas y caminos seguros, con avances, retrocesos, perplejidades y logros, debían dominar nuestra presentación.

Además si la comunidad científica internacional, a pesar de la desigualdad de contextos y niveles, es de una relativa homogeneidad, en modos y objetivos, en la gran mayoría de los sectores, el de las industrias de la lengua es una excepción por la extrema variedad de los actores que en él intervienen: observatorios industriales, instituciones de coordinación de las políticas de los estados, entidades de financiación, grandes multinacionales, pequeñas empresas con una fuerte determinación innovadora, organismos nacionales e internacionales interesados en la defensa y promoción de sus lenguas, etc. Esta proliferación de instancias viene acompañada de una amplia gama de estrategias de asentamiento y expansión: hay quienes fabrican productos que funcionan de verdad, hay quienes no pasan, voluntaria o involuntariamente, del intento, y los hay que escogen el simulacro convencidos de la capacidad promotora del teatro tecnológico; hay quienes financian la investigación y la experimentación y hay quienes se agitan buscando y anunciando fuentes de financiación que no acaban de aparecer; hay quienes comienzan a estructurar este campo mediante normas generales, convenios profesionales y pautas industriales comunes y hay quienes se reúnen, peroran, discuten sin conseguir, aparentemente, ningún resultado.

Y digo, aparentemente, porque en todos los períodos iniciales, sea cual sea el ámbito, es imposible predecir qué vías acabarán siendo las más efectivas. En más de una ocasión el anuncio prematuro de un descubrimiento científico aún no del todo logrado o de un producto industrial aún no disponible han espoleado la capacidad creadora de quienes lo anunciaban y de sus competidores y

transformando en realidad lo que era sólo tentativas cuando no simples previsiones. De igual manera, en las reuniones entre expertos y especialistas de este sector, aparentemente tan estériles, se intercambian competencias, ideas, experiencias que, en ocasiones, contribuyen, de forma decisiva, a hacer avanzar el proceso creador. Por eso he incluido la contribución de Erik Lambert en esta parte, porque considero que el anuncio prematuro de la máquina de escribir con entrada vocal, en base a los intentos de IBM, Kurzweil o Speech Systems a que procede dicho autor, tiene, a su manera, tanto interés como la presentación, también recogida entre estas muestras industriales, que Pierre Le Loarer y Bernard Normier hacen de las actividades y productos —en especial SAPHIR y ALEXIS— de la sociedad ERLI que dirigen, que es una joven empresa europea de gran dinamismo.

Por la misma razón se han incorporado los textos de Robert Mahl y Philippe Amiel que nos ayudan a comprender cómo y por qué los estados y las grandes empresas pueden lanzar proyectos muy ambiciosos que después de un inicio prometedor acaban desapareciendo sin haber logrado los objetivos que se proponían. El Proyecto Nacional de Traducción Asistida por Ordenador (TAO) de que nos habla el primero; y el Proyecto Eurolexic, promovido por cuatro grandes empresas editoriales europeas movilizadas por el grupo Hachette, a que se refiere el segundo, son excelentes ilustraciones del carácter experimental y de ensayo que tienen estas primeras tentativas y de la capacidad aleccionadora de sus logros y de sus malogros. La verificación y corrección ortográfica automatizada es una de las aplicaciones industriales que tienen mayor aceptación y para las que existe un mercado más estabilizado. Jean-François Degrémont, en su presentación, nos ofrece a ese respecto un cuadro de la situación actual, a la par voluntarioso y realista, que no olvida las limitaciones presentes ni excluye las perspectivas de futuro. Los problemas de armonización jurídica de las lenguas europeas, en particular la protección de los intereses de los autores y editores de instrumentos lingüísticos, que aborda en su contribución Christine Poitevin, son un presupuesto esencial para la industrialización de las lenguas, sobre todo en los ámbitos terminológico y lexicográfico y, por dicha razón, me ha parecido necesario que figurase en este lugar.

Después de todo lo dicho pienso que queda clara la opción

central de este libro. No se trata de presentar un conjunto de experiencias concluyentes y de productos a toda prueba, sino más bien de narrar los avatares de un esperanzador proceso en marcha. La naturaleza múltiple y heterogénea de los materiales que forman este Compendio, su diferente nivel de acabamiento y su diversa modalidad de elaboración, su condición incoactiva, que ha llevado a respetar el perfil, en ocasiones un tanto borroso, de toda acción en *status nascenti*, responden a esa opción, así como al propósito concomitante de constituir un marco y un soporte útiles para el desarrollo de las industrias de la lengua en nuestro país. Propósito al que contribuyen también otras acciones e iniciativas en nuestro país e Iberoamérica que hacen esperar que la industrialización del español pueda llevarse a cabo sin que consideraciones inmediatistas e imperativos económicos pongan en peligro su patrimonio cultural o reduzcan su potencia lingüística y literaria.

Este libro es, en todos los sentidos, una obra colectiva. Sin el primer y decisivo impulso de Ernesto García Camarero, verdadero pionero e introductor en la España de los años 60 del estudio de las relaciones entre lengua e informática; sin la competencia y el entusiasmo del Grupo informal de expertos del Consejo de Europa y en particular de los profesores Antonio Zampolli, Maurice Gross, Helmut Schnelle, John Sinclair, Sture Allen y Manuel Alvar Ezquerra; sin los conocimientos y la dedicación casi militante de los profesores Bernard Cassen y Jean-François Degrémont; sin la asistencia incondicional y eficaz de mis colegas en el Consejo de Europa, Gabriele Mazza y Christian Civardi, ni esta obra, ni, sobre todo, el trabajo realizado en favor del patrimonio lingüístico europeo hubieran sido posibles. Quede aquí dicho mi vivo agradecimiento a todos ellos y reivindicados, en exclusiva, los yerros e imperfecciones que los hayan acompañado.